Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 75-82 recibido 04-04-2018 revisado 24-04-2018 aceptado 04-05-2018

Estudio sobre el impacto del corpus de entrenamiento del modelo de lenguaje en las prestaciones de un reconocedor de habla

Study on the impact of the training corpus of the language model on the performance of a speech recognizer

Andrés Piñeiro Martín[1], Carmen García-Mateo[1], Laura Docío-Fernández[1], Xosé Luis Regueira[2]

[1] AtlanTTic Research Center – Escola de Enxeñaría de Telecomunicación – Universidade de Vigo Campus Universitario 36310 Vigo (Spain)

[2] Instituto da Lingua Galega Universidade de Santiago de Compostela Praza da Universidade, 4, 15782 Santiago de Compostela (Spain)

E-mail: apinheiro@gts.uvigo.es, carmen.garcia@uvigo.es, ldocio@gts.uvigo.es, xoseluis.regueira@usc.es

Resumen: Dentro del reconocimiento automático del habla, los modelos de lenguaje estadísticos basados en la probabilidad de secuencia de palabras (n-gramas) suponen uno de los dos pilares sobre los que se basa su correcto funcionamiento. En este trabajo se expone el impacto que tienen sobre las prestaciones de reconocimiento a medida que estos modelos se mejoran con más texto de mejor calidad, cuando estos se ajustan a la aplicación final del sistema, y por lo tanto, cuando se reducen el número de palabras fuera de vocabulario (*Out Of Vocabulary - OOV*). El reconocedor con los distintos modelos de lenguaje ha sido aplicado sobre cortes de audio correspondientes a tres marcos experimentales: oralidad formal, habla en noticiarios, y TED talks en gallego. Los resultados obtenidos muestran claramente una mejora sobre los marcos experimentales propuestos.

Palabras clave: modelos de lenguaje, reconocimiento automático del habla, palabras fuera de vocabulario

Abstract: Within the automatic speech recognition, statistical language models based on the probability of word sequences (n-grams) represent one of the two pillars on which its correct functioning is based. In this paper, the impact they have on the recognition result is exposed as these models are improved with more text of better quality, when these are adjusted to the final application of the system, and therefore, when the number out of vocabulary (*OOV*) words is reduced. The recognizer with the different language models has been applied to audio cuts corresponding to three experimental frames: formal orality, talk on newscasts, and TED talks in Galician. The results obtained clearly show an improvement over the experimental frameworks proposed.

Keywords: language models, automatic speech recognition, Out of vocabulary words

1 Introducción

Hoy en día, los modelos de lenguaje estadísticos (*ML*) son usados en numerosas aplicaciones como el reconocimiento del habla, el reconocimiento de la escritura, el reconocimiento óptico de caracteres (*OCR*), en correcciones ortográficas, etc.

Dentro del reconocimiento automático del habla, los modelos de lenguaje usados definen la estructura de lenguaje, es decir, restringen de forma adecuada las secuencias de unidades lingüísticas más probables. Los más utilizados son los que funcionan como modelos de probabilidad de secuencia de palabras (n-gramas), y son estimados a partir del análisis de grandes cantidades de texto. Estos MLs poseen una integración sencilla con el modelado acústico, pero pueden ser muy generales, requiriendo una adaptación a la tarea de reconocimiento de la que se trate.

En este trabajo se busca analizar el efecto de mejorar el corpus de aprendizaje de los modelos de lenguaje sobre las prestaciones del reconocedor de habla. Para ello, se han reunido amplios corpus de texto de distintas fuentes y con distintas características y temáticas. Los MLs entrenados con estos corpus de textos han sido probados reconociendo cortes de audio dentro de tres marcos experimentales formados

ISSN 1135-5948. DOI 10.26342/2018-61-8

por: cortes de audio de oralidad formal (lecturas literarias y discursos leídos), cortes de audio de telediarios de la TVG, y cortes de audio correspondientes a TED talks en gallego.

A lo largo del trabajo se estudia el efecto de ajustar los MLs en función de la aplicación final del reconocedor, así como la relación entre las palabras fuera de vocabulario de cada uno de los modelos, la WER (*Word Error Rate*) obtenida y la perplejidad.

2 Modelos de lenguaje estadísticos

2.1 N-gramas

Actualmente los modelos de lenguaje más usados en el reconocimiento automático del habla son los probabilísticos basados en n-gramas, los cuales permiten hacer una predicción estadística de la próxima palabra en la secuencia de texto, ya que asumen que la n-ésima palabra depende de las *n-1* anteriores (historia).

Estos modelos describen el lenguaje como cadenas de Márkov de orden *n-1*, donde la probabilidad de que aparezca una palabra depende únicamente de la palabra anterior o palabras anteriores (en función del orden). Por lo tanto, la probabilidad $P(w_1,...,w_m)$ de observar la secuencia de palabras $w_1,...,w_m$ se aproxima como:

$$P(w_1,\ldots,w_m)=\prod_{i=1}^{m} P(w_i|w_1,\ldots,w_{i-1})\approx \approx\prod_{i=1}^{m} P(w_i|w_{i-(n-1)},\ldots,w_{i-1}) \quad (1)$$

Esta última probabilidad condicionada puede calcularse a partir recuentos de frecuencias de la siguiente forma:

$$P(w_i|w_{i-(n-1)},\ldots,w_{i-1})= =\frac{C(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{C(w_{i-(n-1)},\ldots,w_{i-1})} \quad (2)$$

donde $C(w_1,...,w_m)$ es el número de veces que la secuencia $w_1,...,w_m$ ha sido vista en el corpus de entrenamiento (Jurafsky y Martin, 2008).

Los modelos basados en n-gramas poseen una formulación sencilla, facilidad en su implementación y consistencia con los datos de entrenamiento (debido a esto son los más utilizados actualmente), pero también presentan limitaciones:

- El valor de *n* debe ser relativamente pequeño, ya que a medida que este crece aparecen problemas computacionales en la estimación de los parámetros. Debido a esto, los modelos más comunes suelen ser trigramas (n=3) o tetragramas (n=4), aunque comienzan a aparecer pentagramas (n=5) u órdenes mayores.
- Los modelos de n-gramas no se acomodan bien a los cambios en el discurso. Es muy complicado obtener un ML que sea representativo del habla que se quiere reconocer cuando las temáticas del audio varían. Por esto, es necesario actualizar o variar el ML en función del tema o temática del audio que se desea reconocer.
- Estos modelos tienen problemas de dispersión. Existen gran cantidad de eventos u oraciones que no son vistos durante el proceso de entrenamiento y obtienen una frecuencia relativa igual a cero. Esto causa que a frases con alta probabilidad que contengan dicho evento, se les asigne probabilidad cero. Para evitar que el modelo se "rompa" cuando aparecen estos ceros, es necesario aplicar a los modelos suavizados (*smoothing*), donde se asignan probabilidades pequeñas a este tipo de eventos en lugar de probabilidades iguales a cero; e interpolados, donde las probabilidades se combinan con probabilidades de órdenes menores para intentar mantener la información de qué palabras aparecen con una mayor frecuencia en el texto. Un modelo interpolado que combine probabilidades de unigrama (1-grama) y bigrama (2-grama) es definido de la siguiente forma:

$$P_{Interpolated}(w_i|w_{i-1})= =\lambda P(w_i|w_{i-1})+(1-\lambda)P(w_i) \quad (3)$$

Donde $0\leqslant\lambda\leqslant 1$ es el peso del interpolado. Para este caso, cuanto mayor sea, más se parecerá al comportamiento de un modelo bigrama; y cuanto menor sea , más se parecerá al comportamiento de un modelo unigrama (Jurafsky y Martin, 2008).

2.2 Modelos de lenguaje alternativos

Existen modelos alternativos (Vicente et al. 2015) o híbridos que lidian con los anteriores inconvenientes, donde se combina la eficiencia local del modelo de n-gramas con otros que capturan información sintáctica o de larga distancia (información que no puede extraerse de los eventos del modelo de n-gramas).

Los modelos sintácticos usan gramáticas libres de contexto probabilísticas (Probabilistic

Context Free Grammars, PCFGs) para estudiar cómo se relacionan las palabras del corpus, donde el objetivo es basarse más en la gramática. Las PCFGs representan un mecanismo eficiente para modelar las relaciones de larga distancia entre las diferentes unidades léxicas en una oración. Estos modelos de lenguaje alternativos se utilizan en campos como el reconocimiento sintáctico de formas y lingüística computacional, pero no presentan buenos resultados en tareas complejas con grandes vocabularios como el reconocimiento automático del habla, ya que el costo computacional es elevado.

Los modelos factorizados suponen una solución intermedia: poseen la extensión de los basados en n-gramas y son menos demandantes en recursos que los modelos sintácticos. En estos modelos cada palabra *w* es una colección de características o factores, y es posible incorporar conocimiento lingüístico.

Actualmente, con el aumento de la potencia de cómputo disponible, el uso de redes neuronales profundas se ha extendido numerosos campos de aplicación. Dentro del ASR, los modelos de lenguaje basados en redes neuronales recurrentes (*Recurrent neural network langauge models - RNNLM)* están ganando terreno a los modelos de n-gramas clásicos (Mikolov et al. 2011). Sin embargo, estos modelos no pueden ser utilizados de forma sencilla en la decodificación, por lo que la técnica habitual para su aplicación consiste en hacer una etapa de rescoring sobre una decodificación previa que utiliza los clásicos n-gramas. Existen diversos algoritmos que implementan este rescoring como los presentados en (Xu et al. 2018), (Sundermeyer et al. 2014), o (Chen et al. 2017).

2.3 Evaluación de los MLs: Perplejidad

A la hora de evaluar la calidad de un modelo de lenguaje, la perplejidad (*PP*) es la medida típicamente usada.

La perplejidad puede ser considerada como una medida, en promedio, de cuántas palabras diferentes igualmente probables pueden seguir a una palabra determinada, es decir, sobre un conjunto de prueba, es la probabilidad inversa del conjunto, normalizado por el número de palabras. Por lo tanto, para calcularla se necesita tanto un modelo de lenguaje como un texto de prueba. Puede ser calculada de la siguiente forma (Jurafsky y Martin, 2008), para un conjunto de prueba $W = w_1,...,w_m$:

$$PP(W) = \sqrt[m]{\prod_{i=1}^{m} \frac{1}{P(w_i | w_1, \ldots, w_{i-1})}} \qquad (4)$$

Las medidas más bajas de perplejidad representan mejores modelos de lenguaje. Aquí se debe matizar que lo que nos dice la PP es que "modelan mejor el lenguaje" de ese texto de prueba, y no necesariamente que funcionen mejor en los sistemas de reconocimiento de habla.

3 Creación de los MLs

Hoy en día existe una enorme cantidad de texto en formato digital. Si pensamos en el ámbito de Internet, sólo el número libros, artículos, noticias o reportajes que son publicados diariamente y están disponibles al acceso de cualquier usuario es inmenso; pero existen inconvenientes si se quiere usar este texto en el entrenamiento de los modelos de lenguaje. Los principales problemas son:

- La limpieza y el preprocesado de los textos: la variedad de textos disponibles en la red es directamente proporcional a la cantidad de formatos con los que estos están presentes. El unificar formatos y codificaciones a la hora de descargarlos o utilizarlos puede suponer un problema.

- El otro problema principal está relacionado con que los textos utilizados para entrenar los modelos de lenguaje sean representativos del habla que se quiere reconocer. En nuestro caso, con el gallego, no es sencillo conseguir grandes cantidades de texto representativo del idioma, ya que no existen amplios corpus de texto disponibles de forma abierta, el número de diarios o revistas que publican en el idioma son limitados, y los libros que pueden ser encontrados en mayor medida no son representativos de un habla informal o actual del idioma.

Para trabajar sobre los MLs, principalmente para su uso en el reconocimiento de automático de voz y en la traducción automática, la herramienta más utilizada sea probablemente SRI Language Modeling (SRILM), y ha sido la utilizada en nuestro estudio. Esta permite realizar el *n-gram count*, filtrados, interpolados con órdenes inferiores o añadir *smoothing* para evitar romper nuestros modelos de lenguaje. Existen otras opciones como IRST Language Modeling Toolkit (IRSTLM), para trabajar con grandes MLs, o KenLM, pero menos utilizadas.

4 Marco experimental

4.1 Descripción del sistema ASR

Actualmente nuestro reconocedor funciona con cortes de audio a una frecuencia de muestro (f_s) de 16 KHz y en formato *wav (WAVE form audio file format)*. Previo paso al reconocedor, se realiza una segmentación y una parametrización del corte de audio. En esta extracción de parámetros se obtiene los MFCC (*Mel Frequency Cepstral Coefficients*), los cuales son coeficientes para la representación del habla basados en la percepción auditiva humana.

Con respecto al modelado acústico, este utiliza una red neuronal profunda de 5 capas ocultas, con 1024 neuronas y función de activación RELU (Peddinti, Povey y Khudanpur, 2015). La red ha sido entrenada con material correspondiente a TC-STAR (Docío, Cardenal y García, 2006), con 79 horas de horas de habla en castellano; y Transcrigal (García et al., 2004), con 30 horas de habla en gallego.

Por último, el sistema usa modelos de lenguaje de trigramas entrenados con la herramienta SRILM. En el siguiente apartado se explica en mayor detalle cómo se han entrenado.

4.2 Entrenamiento de los MLs

El primer paso para la creación de los modelos de lenguaje ha sido la obtención de texto. Este texto es descargado de Internet y guardado de forma unificada en distintos archivos en función de su origen y características.

Antes del entrenamiento de los modelos, se realiza un primer filtrado con el objetivo de eliminar caracteres extraños, palabras en otros idiomas, errores del proceso de descarga y también de unificar formatos. En este paso la totalidad del texto es divida en oraciones separadas por saltos de línea. Una vez dividido, los signos de puntuación son eliminados (salvo el guión) ya que no aportan información al modelo que utilizará nuestro reconocedor.

El segundo filtrado realizado tiene el objetivo de reducir el tamaño del modelo en función de la probabilidad de que aparezcan las palabras, es decir, se utilizarán sólo las palabras del vocabulario creado que aparecen con una cierta probabilidad.

Una vez se ha preprocesado y filtrado el texto, es posible realizar el proceso de entrenamiento. Los modelos han sido entrenados usando la SRI Language Modeling Toolkit. Se han usado modelos de orden 3, es decir, trigramas. También se ha aplicado un discounting modificado de Kneser-Ney de Chen y Goodman, junto con una interpolación que hace que las estimaciones de probabilidad de orden 3 se interpolen con estimaciones de orden inferior (Stolcke, 2002).

Cada uno de estos modelos entrenados ha sido combinado con un modelo base de trigramas de forma equiprobable. Por último, se realiza la transcripción fonética con la herramienta COTOVIA (Campillo y Rodríguez, 2005) y se crea el grafo que utilizará el reconocedor con las herramientas proporcionadas por KALDI (Povey et al., 2011).

A continuación se explican cada uno de los modelos entrenados, y en la Tabla 1, se observan los principales parámetros de estos modelos. En ella se presenta el número de palabras en vocabulario, el tamaño del texto de entrenamiento, la media de palabras OOV y la perplejidad media sobre los cortes de audio analizados:

ML	Nº palabras en vocabulario	Tamaño del texto de entrenamiento (en millones de palabras)	Porcentaje medio de palabras OOV sobre cortes analizados	Perplejidad media
DOG	210.000	65	9,7 %	789
DUVI	65.000	2,3	11,4 %	616
Wikipedia	450.000	29	6,1 %	703
BEPUB	400.000	317	6,8 %	546
GEV	550.000	81	3,9 %	728
CORGA	420.000	35	3,2 %	582

Tabla 1: Datos sobre los MLs utilizados.

- **DOG**: Modelo entrenado únicamente con texto extraído del DOG (Diario Oficial de Galicia). Se trata de un ML de tamaño medio con texto representativo de un habla muy formal. Aunque este ML posee un tamaño considerable, presenta múltiples tecnicismos o formalismos en su texto, por lo que no se esperan grandes resultados.
- **DUVI**: Modelo entrenado con textos del DUVI (Diario de la Universidad de Vigo). Se trata de un ML pequeño, pero que posee un texto muy limpio y representativo, incluyendo multitud de palabras y expresiones actuales.
- **BEPUB**: Modelo entrenado con un conjunto de 5.000 novelas en castellano traducidas al gallego usando un traductor automático (Alegría et al., 2006). Se debe tener en cuenta que los resultados obtenidos pueden tener errores sistemáticos debidos a la traducción castellano-gallego. El potencial de este modelo consiste, más que en su vocabulario, el cual es cierto que puede no

presentar palabras o expresiones actuales, en la elevada confiabilidad que pueden presentar los n-gramas gracias a la gran cantidad texto sobre el que están entrenados.

- **Wikipedia**: Modelo entrenado con el material en gallego disponible en la Wikipedia. Se trata de un texto que puede reducir el número de palabras fuera de vocabulario, ya que presenta material actualizado y muy variado. Su principal desventaja reside en la falta de texto representativo del habla actual o de diálogos de habla espontánea, ya que está formado en su mayoría por definiciones.
- **Ghoxe + Eroski + Vieiros (GEV)**: Modelo entrenado con texto del diario digital Galicia Hoxe, publicado en Santiago de Compostela; con texto de la página de noticias de Eroski, la cual contiene texto de noticias variadas sobre temas de actualidad; y de Vieiros, otro diario digital editado íntegramente en gallego. Se trata de un corpus extenso, con material limpio y representativo, ya que posee noticias de actualidad de temática variada.
- **CORGA**: Modelo entrenado con textos del Corpus de Referencia del Gallego Actual (CORGA), integrado por distintos textos representativos correspondientes a libros, diarios, revistas, obras de teatro, material audiovisual y blogs. Se trata de un corpus de tamaño medio con texto muy cuidado y muy representativo.

A lo largo del estudio, tras haber realizado las primeras pruebas con los modelos que se acaban de describir, el siguiente paso (fase 2) fue realizar combinaciones de modelos. Las mezclas realizadas han sido las siguientes:

- **Mezcla 1**: uniendo todo el texto de CORGA, GEV y DUVI, y volviendo a entrenar un nuevo ML.
- **Mezcla 2**: combinando directamente los modelos de lenguaje ya entrenados de CORGA, GEV y DUVI en un nuevo ML. De esta forma se busca analizar las diferencias de resultados entre entrenar nuevos modelos con la totalidad del texto o mezclar los modelos ya entrenados.
- **Mezcla 3**: combinando los MLs ya entrenados de CORGA y BEPUB.
- **Mezcla 4**: combinando los MLs ya entrenados de CORGA, GEV, DUVI y BEPUB.

Los modelos DOG y Wikipedia han sido descartados en estas mezclas debido a los malos resultados obtenidos en la fase 1.

En la Tabla 2 se muestra, para cada una de las combinaciones de MLs, el número de palabras en vocabulario, el porcentaje medio de palabras fuera de vocabulario y la perplejidad media sobre los cortes de audio analizados:

ML	Nº palabras en vocabulario	Porcentaje medio de palabras OOV sobre cortes analizados	Perplejidad media
Mezcla 1	720.000	2,5 %	627
Mezcla 2	730.000	2,5 %	608
Mezcla 3	630.000	2,8 %	627
Mezcla 4	900.000	2,3 %	674

Tabla 2: Datos sobre combinaciones de MLs.

4.3 Corpus de análisis.

Las pruebas han sido realizadas sobre tres corpus con características diferentes.

4.3.1 Primer Corpus: Oralidad formal

Corpus con cortes de audio de oralidad formal correspondientes a lecturas literarias y discursos leídos y orales.

Se trata de 30 cortes con una duración media de 3:50 minutos por corte y una duración total de aproximadamente 115 minutos (cerca de 2 horas).

4.3.2 Segundo Corpus: Habla en noticiarios

Corpus con cortes de audio correspondientes a telediarios de la TVG (Televisión de Galicia). Presentan mezclas de habla espontánea y habla planeada y leída, pero con temas y vocabulario más actuales que en el primer corpus.

Está compuesto por 10 audios con una duración media de 34 minutos por corte y una duración total de 340 minutos (5 horas y 40 minutos).

4.3.3 Tercer Corpus: Habla en TED Talks

Corpus con cortes de audio correspondientes a charlas TED Talks en gallego. Estos presentan habla planeada pero no leída, y parte de habla espontánea.

Se trata de 10 cortes con una duración media de 16 minutos por audio y una duración total de 163 minutos (2 horas y 43 minutos).

5 Resultados experimentales

Los siguientes resultados muestran la WER media obtenida con cada uno de los MLs. También se muestra el porcentaje medio de palabras fuera de vocabulario para cada uno de los marcos experimentales en función del ML analizado.

Para los tres marcos experimentales el proceso de análisis y extracción de los resultados será el siguiente: como primer paso (fase 1), se calcula la WER tras reconocimiento y se obtiene el número de palabras fuera de vocabulario usando los MLs “simples”; y como

segundo paso (fase 2), se analiza el efecto de combinar los MLs que mejores resultados han obtenido.

5.1 Resultados con MLs simples – fase 1

En la Tabla 3 se muestra la WER media obtenida en cada uno de los corpus de análisis en función del ML usado, así como el intervalo de confianza (IC) del 95% y la desviación típica.

Los mejores resultados los obtienen los MLs GEV y CORGA. Este último funciona especialmente bien dentro del Corpus 1, donde se consigue reducir en más de un 11 % la WER obtenida por el modelo DOG; mientras que GEV obtiene los mejores resultados en los Corpus 2 y 3, pero reduciendo la WER en menor medida (aproximadamente un 3 % en el Corpus 2 y 4 % en el Corpus 3, ambos con respecto al ML DOG).

WER (%)		DOG	DUVI	WIKI PEDIA	BEPUB	GEV	CORGA
Corpus 1	Valor medio	28,58	28,55	23,85	21,01	20,94	**17,36**
	IC 95%	± 1,98	± 2,16	± 1,83	± 2,01	± 1,84	± 1,84
	Desv. típica	5,50	6,05	5,1	5,63	5,14	5,14
Corpus 2	Valor medio	25,08	24,77	24,1	24,5	**22,13**	23,5
	IC 95%	± 2,34	± 2,3	± 2,34	± 2,17	± 2,19	± 2,17
	Desv. típica	3,27	3,22	3,27	3,03	3,06	3,03
Corpus 3	Valor medio	27,89	26,71	25,64	24,5	**23,57**	24,02
	IC 95%	± 4,76	± 4,46	± 5,09	± 5	± 4,96	± 5,32
	Desv. típica	6,66	6,23	7,11	6,99	6,94	7,44

Tabla 3: Resultados para cada ML.

5.2 Resultados con combinaciones de MLs – fase 2

Si nos fijamos en los resultados obtenidos con las mezclas de modelos presentados en la Tabla 4, para el Corpus 1 se observa que no se consigue reducir la WER media con ninguna combinación, pero sí que se consigue reducir el intervalo de confianza de los resultados. Los mejores resultados de combinaciones para el Corpus 2 son muy similares a los obtenidos con los modelos simples. Sólo para el Corpus 3 se consiguen mejorar ligeramente los resultados obtenidos, por lo que se puede concluir que las combinaciones de modelos no logran reducir de forma significativa la WER media obtenida, y sólo consiguen reducir ligeramente el intervalo de confianza y la desviación típica en los Corpus 1 y 2. La principal ventaja de los modelos combinados reside en que no sería necesario ir cambiando de modelo en función de la temática del audio, ya que presentan unos resultados más robustos en media frente a los tres corpus de estudio que cualquiera de los modelos simples.

WER (%)		Mezcla 1	Mezcla 2	Mezcla 3	Mezcla 4
Corpus 1	Valor medio	17,61	**17,51**	17,55	18,14
	IC 95%	± 1,76	± 1,71	± 1,78	± 1,77
	Desv. típica	4,92	4,78	4,98	4,95
Corpus 2	Valor medio	22,3	**22,22**	24,14	23,57
	IC 95%	± 2,17	± 2,15	± 2,28	± 2,27
	Desv. típica	3,03	3,01	3,19	3,18
Corpus 3	Valor medio	23,35	**23,14**	23,84	23,6
	IC 95%	± 5,32	± 5,26	± 5,19	± 5,21
	Desv. típica	7,44	7,26	7,26	7,28

Tabla 4: Resultados para las combinaciones de MLs.

Es interesante comparar los resultados obtenidos por la Mezcla 1 y la Mezcla 2. En estos se aprecia que la hora de combinar modelos, los valores más bajos de WER se obtienen cuando mezclan los modelos previamente entrenados por separado.

En la Figura 1 se muestra en perspectiva la evolución de la WER media para cada ML en función de cada uno de los Corpus estudiados. En ella podemos observar que la reducción de la WER conseguida en el Corpus 1 es mucho más evidente que para el resto de corpus:

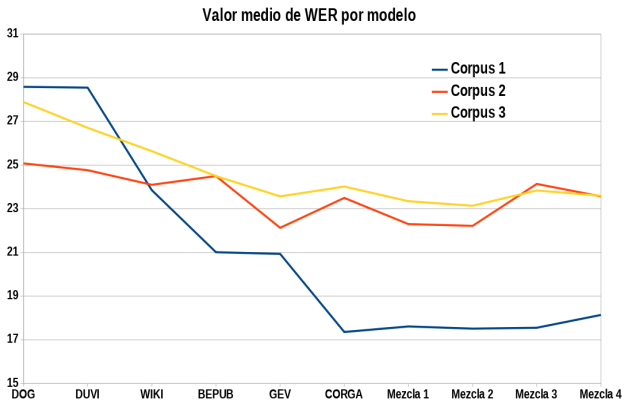

Figura 1: Evolución de la WER media para cada uno de los marcos experimentales.

Fijándonos en el porcentaje medio de palabras OOV, en la Figura 2 se observa su evolución para cada uno de los marcos experimentales. Como era de esperar, los valores más bajos de palabras OOV se obtienen con el modelo que combina más MLs (Mezcla 4). También se aprecia que los valores finales a los que se llega son similares en los tres marcos experimentales, mientras que claramente en el Corpus 1 es en el que la reducción de palabras OOV es mayor.

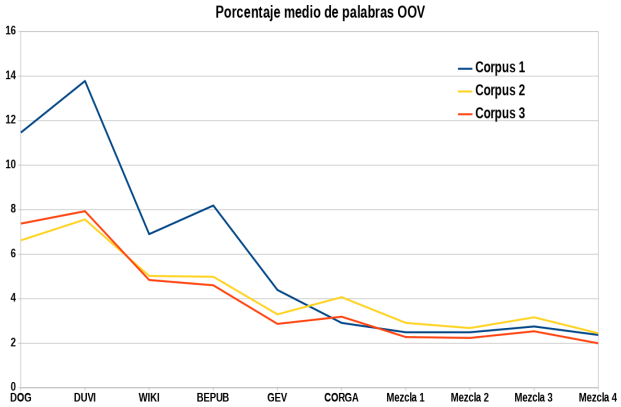

Figura 2: Evolución de las palabras OOV para cada uno de los marcos experimentales.

6 *Discusión*

Si analizamos los resultados obtenidos comparando cómo se consigue reducir el porcentaje de palabras OOV y la WER media obtenida para cada marco experimental, se obtiene lo representado en la Figura 3 (para el Corpus 2). En ella se observa una clara correlación entre el porcentaje de palabras OOV y el valor de WER obtenido, habiendo una progresión muy similar para las dos líneas.

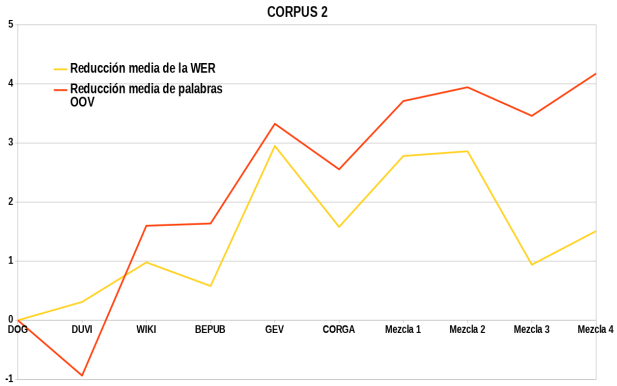

Figura 3: Comparación con DOG de la reducción de la WER y de las palabras OOV para el Corpus 2.

Por otra parte aunque para los tres corpus se observa esta correlación, también se concluye de los resultados obtenidos que reducir el porcentaje de palabras OOV no siempre es sinónimo de reducir la WER obtenida (véanse Figuras 1 y 2).

El comportamiento disimilar entre el Corpus 1 y los otros dos puede obedecer al diferente carácter de las muestras lingüísticas. El Corpus 1 está constituido mayoritariamente por textos leídos (lengua escrita), mientras que el corpus 2 presenta un número elevado de locutores, mezcla heterogénea de tipos de habla (lengua leída, declaraciones de diferentes hablantes, situaciones de ruido, mezcla de gallego y castellano, entre otras). El corpus 3, aunque más homogéneo, está constituido por discurso oral, y cabe hipotetizar sobre si la falta de respuesta de la WER a la suma de MLs (que si tienen efecto en la reducción de OOV, ver Figura 2) está relacionada con que los MLs son fundamentalmente modelos de lengua escrita.

Un análisis más detallado de los errores de reconocimiento puede llevar a reducir todavía más los porcentajes de WER. Una parte de los errores típicos está relacionado con la incorporación de modelos de gallego y castellano, necesarios para el reconocimiento de corpus en que las dos lenguas están presentes (como sucede en el Corpus 2), pero no en los restantes (Corpus 1 y 3). Mas debe tenerse en cuenta que una parte de los errores, al menos en los corpus de lengua oral (no leída), debe asumirse como inevitables, ya que obedecen a dudas y errores de pronunciación de los hablantes, desviaciones de formas, etc., que pueden aparecer marcados como errores, pero en los que el reconocedor en realidad acierta.

7 *Conclusiones y líneas futuras*

La importancia de la calidad de los corpus de entrenamiento con los que se crean los modelos de lenguaje ha quedado reflejada en los resultados presentados.

Se ha conseguido una bajada media de la WER de aproximadamente un 11 % para el corpus de entrenamiento 1, de un 3 % para el corpus 2 y de un 4,75 % para el corpus 3. Por otra parte, se ha comprobado cómo las combinaciones de modelos no logran mejorar de forma significativa los resultados.

También se ha observado que los resultados obtenidos para cada ML dependen claramente del marco experimental sobre el que se está trabajando, es decir, es complicado crear un ML que funcione de forma correcta cuando se varía la temática y características del audio. Como solución a esto, las combinaciones de MLs

obtienen buenos resultados en media frente a los tres corpus de test.

Y por último, la clara correlación (aunque no estrictamente directa) entre el porcentaje de palabras OOV y la WER obtenida evidencia la necesidad de mejorar o adaptar los vocabularios de nuestros modelos.

Como líneas futuras, en recientes estudios se comienza a utilizar las redes neuronales para el entrenamiento de modelos de lenguaje. Aunque se sigue utilizando el entrenamiento "clásico" basado en la predicción estadística como primer paso, se realiza un rescoring de los modelos usando redes neuronales. Los resultados de este rescoring muestran mejoras sobre los resultados experimentales.

Otras posibles líneas de investigación serían aumentar el orden de los n-gramas con los que se entrenan los modelos de lenguaje, puesto que la tecnología actual ya permite trabajar con tetragramas (4-gramas) o incluso pentagramas (5-gramas); o disponer de modelos basados en el habla para poder utilizarlos de forma opcional según el carácter del corpus a reconocer.

Agradecimientos

El trabajo realizado está enmarcado en el proyecto del Plan Nacional TraceThem TEC2015-65345-P y en la red gallega TecAnDaLi ED431D 2016/011 financiada por la Xunta de Galicia. Asimismo se beneficia de las ayudas de la Xunta de Galicia de Grupos de Referencia Competitiva GRC2014/024 y Agrupación Estratéxica Consolidada de Galicia acreditación 2016-2019 y a la Unión Europa a través de los fondos FEDER. Se agradece al Instituto Ramón Piñeiro de la Xunta de Galicia el acuerdo de colaboración para la utilización del material del CORGA y su participación en el etiquetado de los corpus 2 y 3.

Referencias

Peddinti, Vijayaditya, D. Povey y S. Khudanpur. 2015. A time delay natural network architecture for efficient modeling of long temporal context. En *Proceedings of INTERSPEECH*.

Stolcke, Andreas. 2002. SRILM An extensible language modeling toolkit. En *Proceedings of the International Conference on Statistical Language Processing*. Denver, Colorado.

García, Carmen, J. Tirado, L. Docío y A. Cardenal. 2004. Transcrigal: A bilingual system for automatic indexing of broadcast news. *IV International Conference on Language Resources and Evaluation*.

Docío, Laura, A. Cardenal y C. García. 2006. TC-STAR 2006 automatic speech recognition evaluation: The uvigo system. En *Proc. Of TC-STAR Workshop on Speech-to-Speech Translation*. ELRA, París, France.

Jurafsky, Daniel, y J.H. Martin. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

Vicente, Marta, C. Barros, F. Peregrino, F. Agulló y E. Lloret. 2015. La generación de lenguaje natural: análisis del estado actual. Computación y Sistemas. Volumen: 9, n.º 4.

Povey, Daniel, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlícek, Y. Quian, P. Schwarz, J. Silovský, G. Stemmer y K. Veselý. 2011. The Kaldi Speech Recognition Toolkit. En *ASRU Workshop*.

Campillo, Francisco y E. Rodríguez. 2005. Evaluación del modelado acústico y prosódico del sistema de conversión texto-voz Cotovía. En *Procesamiento del Lenguaje Natural. Volumen 35, páginas 5-12*.

Alegría, Iñaki, I. Arantzabal, M. Forcada, X. Gómez, L. Padró, J.R. Pichel y J. Waliño. 2006. OpenTrad: Traducción automática de código abierto para las lenguas del estado Español. En *Procesamiento del Lenguaje Natural*. Volumen: 37, páginas 356-358.

Mikolov, Tomas, S. Kombrink, A. Deoras, L. Bruget y J. Cernocky. 2011. Rnnlm-recurrent neuronal network language modeling toolkit. En *Proc. of ASRU Workshop*.

Xu, Hainan, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey y S. Khudanpur. 2018. A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition. En *ICASSP*.

Sundermeyer, Martin, Z. Tüske, R. Schlüter y H. Ney. 2014. Lattice decoding and rescoring with long-span neural network language models. En *Fifteenth Annual Conference of the International Speech Communication Association*.

Chen, Xie, X. Liu, A. Ragni, Y. Wang y M. Gales. 2017. Future word contexts in neuroal network language models. ArXiv preprint arXiv:170805592.